DOCUMENT RESUME

ED 289 890 TM 870 654

AUTHOR Rengel, Elizabeth
TITLE Agreement between Statistical and Judgmental Item Bias Methods.
PUB DATE Aug 86
NOTE 12p.; Paper presented at the Annual Meeting of the American Psychological Association (Washington, DC, August 22-26, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Aptitude; *Aptitude Tests; Black Students; *Culture Fair Tests; Ethnic Bias; Evaluation Methods; High Schools; Hispanic Americans; Interrater Reliability; *Item Analysis; Racial Bias; Sex Bias; *Statistical Bias; *Test Bias; Test Items; White Students
IDENTIFIERS Ball Aptitude Battery

ABSTRACT

The Ball Aptitude Battery (BAB) was examined for item bias in a sample of 577 high school students in which males and females, as well as three ethnic groups (Blacks, Whites, and Hispanics) were represented. The objectives of the investigation were: (1) to assess the level of interrater agreement for the judgmental method; (2) to find the level of agreement between the judgmental method and several statistical methods; and (3) if agreement was low, to determine whether correspondence was improved when judgmentally biased items were further classified in terms of content or facial validity bias. Two BAB subtests were analyzed for race and sex bias by the judgmental method and by five statistical methods: TID, partial correlation, Rasch difficulty shift, point biserial, and Scheuneman's modified chi square. In the judgmental analysis, the level of interrater agreement among the three expert judges was not significant. The agreements between the statistical and judgmental methods in identification of biased items were also nonsignificant. A final analysis investigated whether agreement between the statistical and judgmental methods was improved when the judgmentally biased items were further classified into one of two types, content validity or face validity bias. Again the results were nonsignificant. (Author/MAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED289890

Agreement Between Statistical and Judgmental Item Bias Methods

by

Elizabeth Rengel
University of Minnesota
Department of Psychology
N218 Elliott Hall
75 East River Road
Minneapolis, MN 55455

Paper presented at the Annual Meeting
of the American Psychological Association
Washington, D.C., August 1986

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Elizabeth Rengel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 870 654

BEST COPY AVAILABLE

Agreement Between Statistical and Judgmental Item Bias Methods

Judgmental analysis at the item level has always been an integral part of the test development process. Within the context of item bias research, however, the judgmental method has taken on the additional task of determining whether the item-to-construct relationship is equivalent across groups. Attention has shifted to include who the judges are, the focus of their judgments, and the systemization of their judgments (Tittle, 1982). The judgmental bias study usually involves a panel of experts (women and/or ethnic minority group members) who review a collection of test items for apparent bias. The focus of the judgmental process becomes the face validity and content validity of the test items for particular subgroups. Face validity in this context refers to the positive representation of a cultural group, while content validity involves the balance, representativeness and appropriateness of the item content relative to the varying educational and life experiences of the cultural groups. Violations of face validity (also known as "facial bias") include demeaning or unfair representation of women and minorities. Items founds biased in terms of content validity involve unequal experience of groups with item format or content: for example, the use of group-specific language.

The issue of most relevance to the present study involves the extent of agreement between the results of judgmental and statistical bias detection techniques. Previous studies have consistently reported low levels of agreement between the methodologies (Jensen, 1977; Plake, 1980; Cotter & Berk, 1981; Sandoval & Miille, 1980). Each of these studies employed an ANOVA-based statistical bias technique and a panel of "experts" (variously defined) for the judgmental analysis.

The objectives of the present study are three-fold: (1) to assess the level of inter-rater agreement for the judgmental method, (2) to investigate the level of agreement between the judgmental method and several statistical methods, and (3) if agreement is low, to determine whether correspondence is improved when judgmentally biased items are further classified in terms of two bias categories, i.e., content or facial validity bias.

The present study differs from previous agreement studies in several respects. First, it formally investigates the level of agreement among the expert judges. Second, it uses a variety of statistical bias techniques, including an IRT-based method. Finally, it investigates how further classification of judgmentally biased items affects agreement with the statistical methods.

METHOD

Instrumentation. Two tests were selected from the sixteen-test Ball Aptitude Battery (BAB) according to the criteria described by Rengel (1986). They are described briefly as follows:

1. Inductive Reasoning (30 items). Measures the ability to identify a rule; to reason from specific information to a general principle. The task involves the identification of three pictures from a series of six that share a common element.

2. Vocabulary (80 items). Measures knowledge of English vocabulary. The examinee is asked to select from among five choices the word most equivalent in meaning to the test word.

The first test (except for the instructions) is pictorial in content, while the second is entirely verbal.

Subjects. The entire BAB was administered in 1977-78 to a sample of 1006 students from 36 high schools in Illinois and Texas. A subsample of 577 students was selected to provide approximately equal representation of both genders and three ethnic groups (Black, Hispanic and White). Each sex-ethnic group was chosen so that all groups were comparable in distribution of socio-economic status. Within each sex-ethnic group, approximately 49% were in the lower SES group, 33% in the middle group, and 18% in the upper group. Students' ages ranged from 15 to 19 years, with a mean age of 17.53 and a standard deviation of 0.79.

Statistical bias analyses. Each test was analyzed for race and sex bias by five statistical item bias techniques:

1. Transformed item difficulty (TID: Angoff & Ford, 1973). Bias is defined as item by group interaction in the ANOVA sense of the term.

2. Modified chi square (C2; Scheuneman, 1979). Item bias is defined is terms of differential probabilities of a correct response for test-takers from different groups of the same ability level.

3. Rasch item difficulty shift (Draba, 1978). Bias is indicated by different within-group item difficulties as estimated by the one-parameter Rasch model.

4. Partial correlation (R; Stricker, 1982). A biased item displays a significant correlation with group membership after partialling out total test score.

5. Point biserial method (PBIS; modified from Merz & Grossen, 1979). A biased item displays significantly different item-total correlations across groups.

The statistical analyses were performed on each pairwise race or sex group comparison (i.e., Black-White, White-Hispanic, male-female) for a total of 30 statistical bias analyses (2 tests x 5 methods x 3 group comparisons). Each item was classified as "biased" or "unbiased" according to the bias criterion associated with a particular bias method.

Judgmental bias analysis. The judgmental method involved the qualitative evaluation of individual test items for race or sex bias. Three judges (one White woman and two ethnic minority men) were selected for their backgrounds in psychometrics and their sensitivity to race and sex bias issues. In the instructions to the judges, item bias was defined as whatever distorts the

meaning of item performance for a group and results in the mispresentation of the group's abilities. Five categories of item bias were defined, adapted from Tittle (1982):

1. Differential distractor strength (e.g., more than one correct answer for a group).
2. Language offensive to a group.
3. Differential meaning or interpretation of words or situations in the item.
4. Unfair representation of a group's lifestyle.
5. DIfferential familiarity with situations presented in the item.

Three of the bias categories (1, 3, 5) were designed to correspond to violations of content validity, while the others (2, 4) related to issues of face validity or facial bias.

Each judge was given test copies keyed with the correct responses. In the space after each item, the judge indicated whether, in his or her judgment, the item was biased against one or more of the sex or race groups. The number of items identified as biased was left to the individual judge's discretion. After identifying a biased item, the judges were asked to explain briefly their bias decision in terms of the five bias categories described above. The judges' bias decisions were made independently of one another and were based entirely on item content and format; no item statistics were provided.

RESULTS

Judgmental bias analysis. The number of biased items across the two tests varied by judge, ranging from six to 26. All three judges found considerably more items biased against ethnic minorities or females than against Whites or males (56 vs. 1 and 12 vs. 1, respectively). Furthermore, the White female judge identified more instances of sex bias than race bias (5 items vs. 4 items), while the ethnic minority men judges identified more cases of race bias than sex bias (41 vs. 1 and 12 vs. 8).

Overall inter-rater agreement was assessed using Fleiss' (1971) generalization of Kappa. The results are presented in Table 1. None of the resulting values differed significantly from zero at the .05 level, indicating that agreement among the three judges was at the chance level. There was a nonsignificant trend toward higher levels of agreement in the identification of sex-biased than race-biased items.

The task of classifying items as biased or unbiased by the judgmental method was complicated by the low inter-rater agreement. It would have been preferable to classify biased items by a consensus of at least two of the three judges, but since such agreement was rare (only two instances), it was decided to consider an item judgmentally biased whenever it had been so classified by one or more judges. The number of biased items identified by this decision rule was, for Inductive Reasoning: gender comparison, 6; Black-White, 5; White-Hispanic, 9; for

Vocabulary, gender, 7; Black-White, 22; White-Hispanic, 20.

Statistical bias analysis. The distribution of biased items by test, bias method and group comparison was presented in Rengel (1986); the agreement among statistical bias methods was described by Rengel, Sung, and Dong (1985).

Agreement between statistical and judgmental bias methods. The agreement between the statistical and judgmental methods in identification of biased items was evaluated by Cohen's (1960) Kappa coefficient of agreement between two judges (or in this case, two bias methods). The results are presented in Table 2 for the BAB tests separately and combined. None of the resulting Kappa's differed significantly from zero at the .05 level, indicating that agreement among bias methods was no better than would be expected by chance.

Agreement tended to be slightly higher for Inductive Reasoning than for Vocabulary, and for the gender comparison than the race group comparisons. It was difficult to discern trends with respect to agreement with specific statistical bias methods. With notable exceptions, agreement tended to be higher for the TID, Rasch, and C2 methods than for the partial correlation or PBIS methods.

Further classification of judgmentally biased items. It has been suggested that the lack of agreement between the statistical and judgmental methods may be due in part to the failure of the latter to distinguish between items biased in terms of content and face validity. The judgmental analysis, it will be recalled, required the judges to classify biased items in terms of violations of the two bias types. Perhaps the statistical bias methods are more sensitive to one type of bias than the other, in which case low levels of agreement are to be expected. In the final data analysis, agreements between the statistical and judgmental bias methods were evaluated separately for the content and face validity biased items.

The majority of the judgmentally biased items for the two tests (60 of 69, or 87%) were found to violate content validity. The distribution by bias type, group comparison and BAB test were provided by Rengel (1986). When the agreement between the statistical methods and the two new judgmental classifications was assessed, none of the resulting Kappa's attained statistical significance. (It should be mentioned that the skewed distributions of content and face validity biased items resulted in markedly disparate marginals in the contingency tables for the latter, and Kappa values were necessarily very low.) Several nonsignificant trends were discerned, however: agreement was generally higher for Inductive Reasoning than Vocabulary, and for sex-biased than race-biased items. When there was a difference in agreement between the judgmentally biased items classified in terms of face validity and content validity, the latter was usually the higher of the two. There was little to distinguish

among the statistical bias methods in terms of their agreement with either judgmental method.

DISCUSSION

The lack of agreement between the judgmental and statistical bias methods was probably the result of a combination of factors. Reynolds (1982) has noted that it would be unrealistic to expect the "armchair" or expert minority panel approach to predict biased items accurately when researchers have had so little success in detecting common characteristics among items identified by the presumably more objective statistical methods. Shepard (1982) has added that at least one study (Sandoval & Miille, 1980) did not use judges who could be considered "experts" in the ability domain under consideration. Flaugher (1978), commenting on the "futility" of the judgmental approach, has noted that agreement among judges, when reported, is invariably very low, and correlations with an unreliable index are necessarily suspect. Perhaps the most important factor accounting for the discrepancy is the multi-faceted nature of the concept of bias iteself. Flaugher (1978) has described a variety of aspects of test bias, in which he included differences in test scores, content validity bias, differential validity, and "atmospheric" bias. Flaugher concluded "[I]t seems essential to keep in mind all of these various aspects of the definition of test bias . . . [L]et us not be confused about whether one of these aspects is the <u>real</u> issue of test bias. As it happens, they are <u>all</u> the real issue (p. 678). When bias is defined so inclusively, perhaps it is not surprising that the various bias methods fail to agree. When item bias is explicitly defined and manipulated, as in the studies by Ironson and her associates, agreement is greatly improved, approaching or exceeding the level of agreement among the statistical bias methods themselves.

Table 1

Overall Inter-Rater Agreement in Identification of Biased Items, by BAB Test and Group Comparison

Group comparison	Kappa	Standard error (k)	$\frac{k}{SE\ (k)}$
	Inductive Reasoning		
Gender	.282	0.383	0.736
Black-White	-.047	0.421	-0.112
White-Hispanic	-.111	0.301	-0.369
	Vocabulary		
Gender	.143	0.402	0.356
Black-White	-.087	0.198	-0.439
White-Hispanic	-.099	0.206	0.481
	Combined tests		
Gender	.116	0.279	0.416
Black-White	-.082	0.181	-0.453
White-Hispanic	-.093	0.169	-0.550

Table 2

Kappa Coefficients of Agreement Between Judgmental and Statistical Bias Classifications, by BAB Test and Group Comparison

Group comparison	Statistical bias methods					Median kappa
	TID	R	PBIS	Rasch	C2	
Inductive Reasoning (30 items)						
Gender	.250 (25)	.348 (30)	-.223 (30)	.432 (30)	.296 (30)	.296
Black-White	.273 (24)	.130 (30)	.296 (30)	.379 (30)	.366 (30)	.296
White-Hispanic	-.221 (24)	.216 (30)	-.064 (30)	.083 (30)	.000 (29)	.000
Vocabulary (80 items)						
Gender	.243 (80)	.100 (80)	-.078 (80)	.219 (80)	.142 (78)	.142
Black-White	-.003 (80)	-.037 (80)	.123 (80)	-.059 (80)	.158 (55)	-.003
White-Hispanic	.200 (80)	-.031 (80)	.081 (80)	.071 (80)	.123 (55)	.081
Combined tests (110 items)						
Gender	.246 (105)	.139 (110)	-.111 (110)	.268 (110)	.199 (108)	.199
Black-White	.052 (104)	-.008 (110)	.172 (110)	.045 (110)	.212 (83)	.052
White-Hispanic	.092 (104)	.019 (110)	.032 (110)	.095 (110)	.075 (84)	.075

Note. Numbers in parentheses denote number of items in analysis.

Table 3

Kappa Coefficients of Agreement Between Statistical Item Bias Methods and Judged Facial or Content Biased Items, by Group Comparison

Method	Inductive Reasoning		Vocabulary	
	Face validity	Content validity	Face validity	Content validity
		Gender		
TID	-.071 (25)	.324 (25)	.146 (80)	.128 (80)
R	-.062 (30)	.424 (30)	.045 (80)	.065 (80)
PBIS	-.060 (30)	-.190 (30)	-.078 (80)	-.027 (80)
Rasch	-.052 (30)	.512 (30)	.126 (80)	.120 (80)
C2	-.050 (30)	.364 (30)	-.065 (78)	.397 (80)
		Black-White		
TID	.000 (24)	.272 (24)	-.025 (80)	.015 (80)
R	.000 (30)	.188 (30)	-.017 (80)	-.026 (80)
PBIS	.000 (30)	.363 (30)	-.025 (80)	.142 (80)
Rasch	.000 (30)	.451 (30)	-.024 (80)	-.046 (80)
C2	.000 (28)	.015 (28)	-.034 (80)	.180 (55)
		White-Hispanic		
TID	.000 (24)	-.221 (24)	.046 (80)	.214 (80)
R	.000 (30)	.215 (30)	.038 (80)	-.027 (80)
PBIS	.000 (30)	-.064 (30)	.037 (80)	.094 (80)
Rasch	.000 (30)	.083 (30)	-.070 (80)	.062 (80)
C2	.000 (29)	.000 (29)	1.000 (55)	-.034 (55)

Note. Numbers in parentheses denote number of items in analysis. None of the Kappa's were significant at the .05 level.

References

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Cotter, D.E., & Berk, R.A. (1981, April). Item bias in the WISC-R using black, white, and hispanic learning disabled children. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA. (ERIC Document Reproduction No. ED 206 631)

Dohm, T.E., & Sung, Y.H. (1980, June). Concurrent, differential, and construct validities of the Ball Aptitude Test. (Technical Report No. 5). Glen Ellyn, Illinois: Ball Foundation.

Draba, R.E. (1978). The Rasch model and legal criteria of a "reasonable" classification. Unpublished doctoral dissertation, University of Chicago.

Flaugher, R.L. (1978). The many definitions of test bias. American Psychologist, 33, 671-679.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382.

Ironson, G.H., & Craig, R. (1982). Item bias techniques when amount of bias is varied and score differences between groups are present. Final Report NIE-G-81-0045. Tampa: University of South Florida. (ERIC Document Reproduction No. ED 227 146)

Jensen, A.R. (1977). An examination of culture bias in the Wonderlic Personnel Test. Intelligence, 1, 51-64.

Merz, W.R., & Grossen, N.E. (1979). An empirical investigation of six methods for examining test item bias (Final report NIE-6-73-0067). Sacramento State University Foundation, CA. (ERIC Document Reproduction No. ED 178 566)

Plake, B.S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: one step in the test validation process. Educational and Psychological Measurement, 40, 397-404.

Rengel, E., Sung, Y.H., & Dong, H.K. (1985, August). Effect of deletion of biased test items on predictive validity. Presented at the meeting of the American Psychological Association, Los Angeles, CA.

Rengel, E. (1986). [Effect of deletion of biased test items on test reliability and validity]. Unpublished manuscript.

Reynolds, C.R. (1982). The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), Handbook of school psychology (pp. 178-208). New York: Wiley.

Sandoval, J., & Miille, M.P.W. (1980). Accuracy of judgment of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

Scheuneman, J. (1979). A method for assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L.A. (1982). Definitions of bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 9-30). Baltimore: Johns Hopkins.

Stricker, L.J. (1982). Identifying test items that perform differentially in population subgroups: a partial correlation index. Applied Psychological Measurement, 6, 261-273.

Subkoviak, M.J., Mack, J.S., Ironson, G.H., & Craig, R.D. (1984). Empirical comparison of selected item bias detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Sung, Y.H., & Dawis, R.V. (1981). Technical manual: Ball Aptitude Battery. Glen Ellyn, Illinois: Ball Foundation.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore: Johns Hopkins.

Abstract

This study investigated the agreement between the judgmental and statistical bias techniques in the identification of biased items. The sample consisted of 577 high school students in which both genders and three ethnic groups (Blacks, Whites, and Hispanics) were equally represented. Two aptitude tests were analyzed for race and sex bias by the judgmental method and by five statistical methods: TID, partial correlation, Rasch difficulty shift, point biserial, and Scheuneman's modified chi square.

The results of the statistical bias analyses were reported by Rengel, Dong, and Sung (1985). In the judgmental analysis, the level of inter-rater agreement among the three expert judges was not significant at the .05 level. The agreements between the statistical and judgmental methods in identification of biased items (as indexed by Kappa) were also non-significant. A final analysis investigated whether agreement between the statistical and judgmental methods was improved when the judgmentally biased items were further classified into one of two bias types, content- or face-validity bias. Again the results were nonsignificant.